UNIVERSITE BADJI MOKHTAR-ANNABA

DEPARTEMENT D'INFORMATIQUE
M2 SID

# Prétraitement de Données collectées via le web scraping

*par*

**Dr. Samira LAGRINI**

Année universitaire:2025/2026

# Introduction

- Les données collectées via le web scraping sont souvent brutes, non structurées et incomplètes.
- Ces données ne peuvent pas être exploités directement pour l'analyse, l'entraînement de modèles ou même stocker dans des bases de données

Une étape de prétraitement est nécessaire.

# Flux de Données dans le Web Mining

# Pourquoi le prétraitement des données est-il important

Les données de web scraping sont collectées depuis de nombreux sites, cela entraîne les problèmes :

- Formats incohérents : les dates, les chiffres et les textes peuvent être formatés différemment d'un site à l'autre.
- Données manquantes.
- Données en double.
- Données supplémentaires non pertinentes (ex. des publicités)
- Les données collectées peuvent contenir des balises HTML et du code indésirables.

# Les Conséquences Invisibles des Données Non Prétraitées

Les données non prétraitées entraînent :

Des décisions erronées

Des analyses incorrectes et des interprétations erronées

Des revenus perdus à cause de campagnes marketing inefficaces

Des modèles biaisés

Une réputation endommagée (paraître peu fiable)

Un temps perdu à corriger des erreurs au lieu d'analyser

# Étapes de prétraitement de données extraites

# Nettoyage des données

## ➢Suppression des doublons (entrées dupliquées)

Les doublons faussent l'analyse en exagérant la fréquence d'un élément.

## ➢ Gestion des valeurs manquantes

Certaines données peuvent être incomplètes. Il faut décider soi de supprimez les lignes ou colonnes avec des données manquantes ou remplissez les valeurs manquantes avec des estimations (imputation : utilisez la moyenne, la médiane ou la valeur la plus fréquente.)

# Standardisation des formats

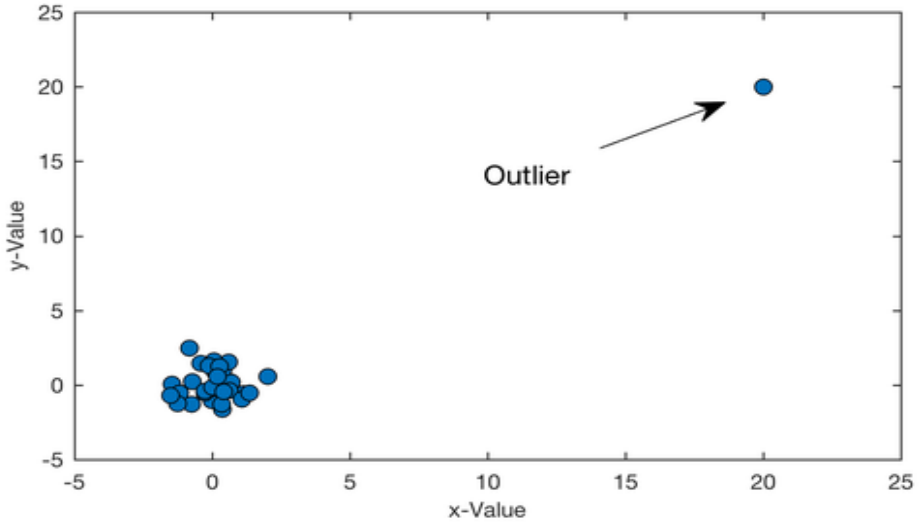
- ✓ Dates : Convertissez toutes les dates dans un format unique (MM/JJ/AAAA vs. JJ/MM/AAAA).
- ✓ Nombres : Utilisez le même séparateur décimal et séparateur de milliers.
- ✓ Texte : Convertissez tout le texte en minuscules ou majuscules.
- ✓ Supprimez les espaces supplémentaires.
- ✓ Unités : Toutes les mesures doivent être dans les mêmes unités (ex : "€ " au lieu de "$")

# Détection et gestion des valeurs aberrantes (Trouver les éléments hors norme)

- Les valeurs aberrantes sont des données extrêmes qui diffèrent du reste des données et peuvent être des erreurs ou des points inhabituels.

**Comment ????**

✓ Détection par visualisation graphique.




✓ Méthodes statistiques : Calculer **scores z** pour identifier les valeurs éloignées de la moyenne.

$$z = \frac{X - \mu}{\sigma}$$

Où :

- $X$ est la valeur observée.
- $\mu$ est la moyenne des données.
- $\sigma$ est l'écart-type des données.

# Normalisation des données (Mettre les données sur la même échelle)

- Ajuster les valeurs de manière uniforme, en particulier pour les variables numériques.
- Utile pour comparer des ensembles de données provenant de différentes échelles ou unités.

**Comment????**

✓ **Min-Max Scaling**
Redimensionner les données pour qu'elles se situent dans un intervalle [0, 1]

$$X_{\text{normalisé}} = \frac{X - \min(X)}{\max(X) - \min(X)}$$

✓ **le score Z**

$$z = \frac{X - \mu}{\sigma}$$

✓ Le **décimal scaling**
Diviser les valeurs par 10, 100, 1000, etc., pour réduire l'échelle des données

# Transformation des données

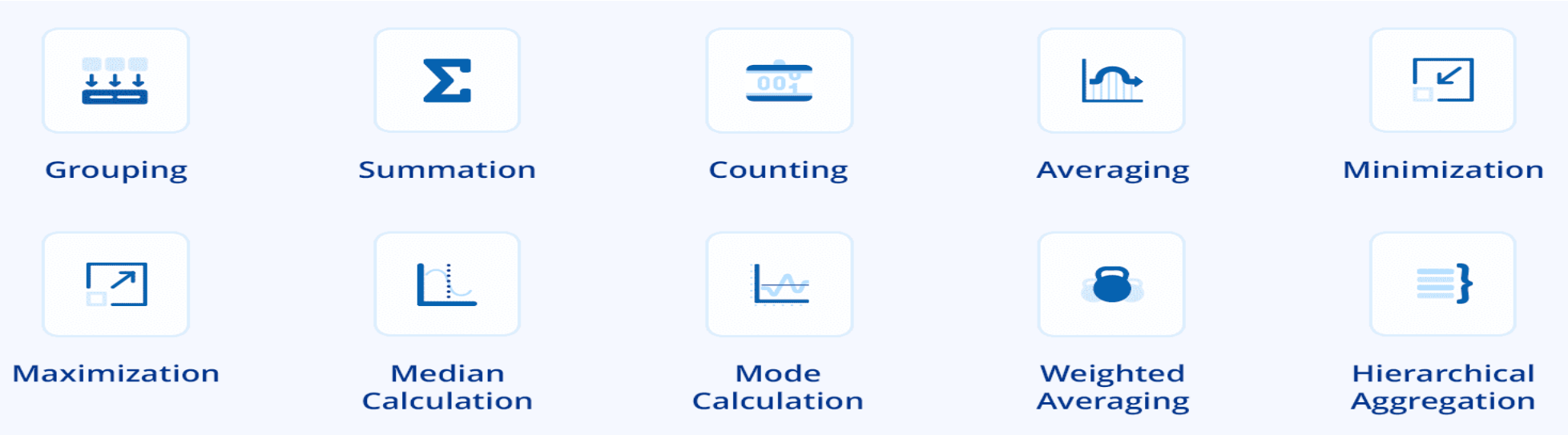
- Parfois, il est nécessaire de modifier la structure de vos données pour faciliter leur analyse.
- Cela permet de rendre les données plus accessibles et adaptées aux différents types de modèles et d'analyses.

**Comment**

# Agrégation

- Combinez des données en résumés (ex, calculer les ventes totales/mois ou les moyennes par catégorie).
- Permet de réduire la complexité des données et de mettre en évidence les tendances globales.

Techniques:



# Encodage des variables catégorielles

- Si les données comprennent des variables qualitatives (comme les couleurs, les catégories de produits, etc.), elles doivent être converties en un format numérique.

*Exemple : Si les données contiennent des informations sur des catégories de livres (fiction, non-fiction, science-fiction), il faut encoder ces catégories sous forme de nombres (par exemple, fiction = 1, non-fiction = 2).*

# Filtrage des données

- **Sélection des colonnes ou des attributs nécessaires** : lors de scraping de nombreuses informations sont collectées, mais souvent seules certaines colonnes sont pertinentes pour l'analyse.

*Exemple : Si vous avez scrappé une page de critiques de films, mais que vous ne vous intéressez qu'à la note et au texte de la critique, il est possible d'ignorer (de filtrer) les autres colonnes comme l'ID de l'utilisateur ou la date de la critique.*

# Enrichissement des données

- **Ajout de données externes** : Parfois, il est possible de compléter les données avec des informations provenant d'autres sources.

*Exemple : Si on a scrappé des informations sur des livres (titre, auteur, critique), il est possible d'ajouter des informations sur les prix des livres en les scrappant à partir d'un autre site.*

- **Création de nouvelles colonnes** : il est possible de dériver de nouvelles informations à partir des données existantes.

*Exemple : À partir d'une colonne de texte de critique, on peux créer une nouvelle colonne indiquant si la critique est positive ou négative, en appliquant une analyse de sentiment.*

# Outils de nettoyage des données

- Des bibliothèques classiques et plateformes permettent le prétraitement des données à l'aide de scripts.
- Nécessite des compétences en programmation, mais offrent un contrôle total sur le processus de prétraitement.

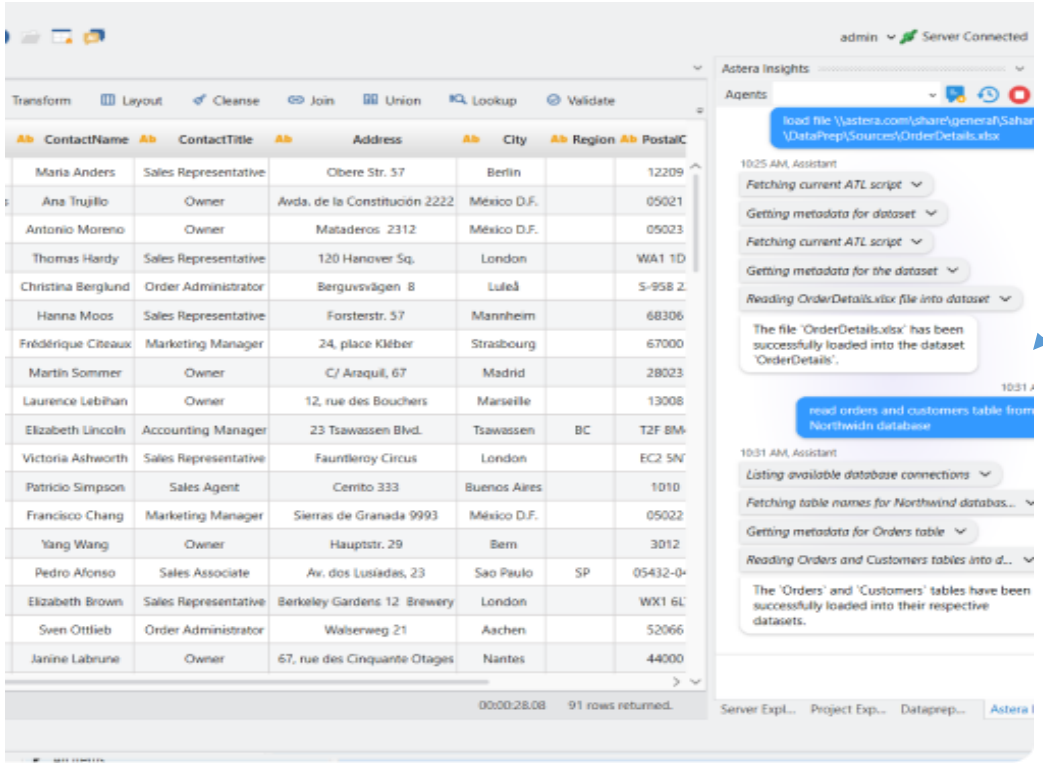
***Exemple :***

- **Python ( avec Pandas library)**

**OpenRefine** : Logiciel open-source pour nettoyer des données complexes.

# Outil récente basé sur l'IA conversationnelle

- Utilise des modèles d‘IA pour rendre le prétraitement plus intuitif et automatisé.




La plateforme **Astera DataPrep** lancée en **2025** offre une interface conversationnelle basée sur l'IA pour prétraiter les données sans nécessiter de compétences en programmation.

*Exemple:* Nettoyage des données avec Python

```
import pandas as pd
# Load  scraped data ( a CSV file)
data = pd.read_csv(“scraped_data.csv”)
# Removing Duplicates
data.drop_duplicates(subset=[“product_id”], keep=”first”, inplace=True)
#  Handling Missing Values (replace with average price)
average_price = data[“price”].mean()
data[“price”].fillna(average_price, inplace=True)
#  Standardizing Formats (convert dates to YYYY-MM-DD)
data[“date”] = pd.to_datetime(data[“date”]).dt.strftime(‘%Y-%m-%d’)
# Detecting and Managing Outliers (remove prices above a threshold)
data = data[data[“price”] < 1000]
# Data Transformation (create a new column for price per unit)
data[“price_per_unit”] = data[“price”] / data[“quantity”]
# Save the cleaned data
data.to_csv(“cleaned_data.csv”, index=False)
print(data.head())
```

# Conclusion

- Le prétraitement des données est une étape fondamentale pour transformer des données brutes en informations exploitables.
- En nettoyant, transformant, et structurant les données de manière adéquate, on garantisse qu'elles seront prêtes à être stockées dans des bases de données complexes et utilisées dans des modèles analytiques ou prédictifs.

# Travaux pratiques

# TP 2 : Prétraitement des Données Collectées via Web Scraping

## Énoncé :

Appliquer les techniques de prétraitement sur les données déjà collectées via un processus de web scraping dans le TP précédent en utilisant Python, OpenRefine et Astera DataPrep et comparer les résultas.